基于动态隧道系统的 K-means 聚类算法研究*

吕 佳
(重庆师范大学 数学与计算机科学学院,运筹学与系统工程重庆市市级重点实验室,重庆 400047)

摘要:针对 K-means 聚类算法易陷入局部极小的问题,利用动态隧道算法在解决全局最优化问题中的有效性,将算法中的动态隧道过程引入到 K-means 聚类算法中,提出了一种基于动态隧道算法的 K-means 聚类算法。该算法在 K-means 聚类算法寻优得到的局部极小值基础上,利用动态隧道过程寻找更小的能量盆地,再将其值提交给 K-means 聚类算法进行迭代寻优,重复该过程,直到找到全局最小值。理论分析和仿真实验证明,该算法的聚类效果要优于 K-means 聚类算法。

关键词:K-means 聚类算法;全局最优化;目标函数;动态隧道系统;能量盆地
中图分类号:TP181 **文献标识码**:A **文章编号**:1672-6693(2009)01-0073-05

K-means 聚类算法是一种经典的基于划分的硬聚类算法,该算法因其思想简单易行,时间复杂性接近线性,对大规模数据的挖掘具有高效性和可伸缩性而被广泛应用在模式识别、图像处理、特征提取、故障诊断中。但由于该算法需预先确定聚类数目 K 且对初值敏感以及算法易陷入局部极小等使其应用存在着一定的局限性。文献[1]中 Bradley 的优化算法采用多次取样数据集两次聚类以获取最优初值的思想,有效地解决了 K-means 聚类算法对初始值的选择具有较大依赖性的问题。Brown and Huntley 于 1991 年提出了用模拟退火法来获取合并最优值。文献[2,3]引入了模糊集和核的思想来解决 K-means 聚类算法不能进行软划分的问题。文献[4-6]分别采用遗传算法和免疫算法改进 K-means 聚类算法。

K-means 聚类算法实质是寻找一组中心矢量,使各样本到中心矢量的距离平方和达到最小,它本质上就是一个全局最优化求解问题,采用了所谓的爬山技术来寻找最优解,因此容易陷入局部极小值点。在目前的全局最优化问题的研究中,提出了一种动态隧道算法[7]。该算法重复以下的两个过程,一是动态优化过程,在该过程中找到一个局部的最小点,二是动态隧道过程,该过程以该局部最小点为初始值找到一个更小的能量盆地,即找到一个新的起始点来提供给动态优化过程。这两个过程交替进行,直到在动态隧道阶段找不到更小的能量盆地为止。因此,为了有效避免 K-means 聚类算法易陷入局部极小的缺陷,笔者将动态隧道过程引入到 K-means 聚类算法中,提出一种基于动态隧道系统的 K-means 聚类算法,使其能在 K-means 聚类算法找到极小值后再利用动态隧道系统寻找更小的能量盆地,从而找到目标函数的全局最优点。

1 K-means 聚类算法

K-means 聚类算法目标是根据输入参数 k,将数据集划分成 k 个簇。算法首先随机选取 k 个点作为初始聚类中心,然后计算各个样本到聚类中心的距离,把样本归到离它最近的那个聚类中心所在的类;对调整后的新类计算新的聚类中心,如果相邻两次的聚类中心没有任何变化,说明样本调整结束,聚类准则函数 J_c 已经收敛。

该算法属于动态聚类法,其迭代过程采用按批修改方法,即在每次迭代中都要考察每个样本的分类是否正确,若不正确,就要调整。在全部样本调整完后,再修改聚类中心,进入下一次迭代。如果在一次迭代算法中,所有的样本被正确分类,则不会有调整,聚类中心也不会有任何变化,这标志着 J_c 已经收敛,算法结束。该算法流程如下[7]。

Step1 给定数据规模为 n 的数据集,令 $I = 1$,选取 k 个初始聚类中心 $V_j(I)$,$j = 1,2,3,\cdots,k$;

* 收稿日期:2008-10-10
资助项目:国家自然科学基金(No. 10171118);重庆市教委科学技术研究项目(No. KJ060818);运筹学与系统工程重庆市市级重点实验室开放课题(No. YC200804)
作者简介:吕佳,女,讲师,博士研究生,研究方向为数据挖掘与最优化技术。

Step2 计算每个数据对象与聚类中心的距离，$D(x_i, V_j(I))$，$i = 1,2,3,\cdots,n$，$j = 1,2,3,\cdots,k$，如果满足 $D(x_i, V_k(I)) = \min\{D(x_i, V_j(I)), j = 1,2,3,\cdots,n\}$，则 $x_i \in w_k$；

Step3 计算误差平方和准则函数 J_c

$$J_c(I) = \sum_{j=1}^{k}\sum_{k=1}^{n_j}\| x_k^{(j)} - V_j(I)\|^2 \qquad (1)$$

Step4 若 $|J_c(I) - J_c(I-1)| < \xi$ 则算法结束 否则 $I = I + 1$，计算 k 个新的聚类中心，$V_j(I) = \frac{1}{n}\sum_{i=1}^{n_j} x_i^{(j)}$，$j = 1,2,3,\cdots,k$，返回 Step 2。

K-means 聚类算法是以数据点到原型(类别中心)的某种距离和作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-means 聚类算法以欧式距离作为相似性测度，采用误差平方和准则函数作为聚类准则函数，误差平方和准则函数定义为 $J_c = \sum_{i=1}^{k}\sum_{p\in C_i}\| p - M_i\|^2$。其中，$M_i$ 是类 C_i 中数据对象的均值，p 是类 C_i 中的空间点。它是求对应某一初始聚类中心向量 $V = (V_1, V_2, \cdots, V_k)^{\mathrm{T}}$ 最优分类，使得评价指标 J_c 值最小。

分析误差平方和准则函数发现：K-means 聚类算法是一个最优化求解问题，目标函数存在着许多局部极小点，只有一个是全局最小点。目标函数的搜索方向总是沿着误差平方和准则函数减小的方向进行。不同的初始值使得聚类中心向量 V 沿着不同的路径使目标函数减少。如图 1 所示，目标函数分别沿着 V_A、V_B、V_C 3 种不同的初始值向量的路径逐步减小，分别找到各自对应的最小值。其中，只有 B 点对应的最小值才是全局最小点，而 A、C 两点对应的最小值是局部极小点。这是因为 K-means 聚类算法的目标函数在空间状态是一个非凸函数，非凸函数往往存在很多个局部极小值，只有一个属于全局最小。由于算法每次开始选取的初始聚类中心落入非凸函数曲面的“位置”往往偏离全局最优解的搜索范围，因此通过迭代运算，目标函数常常达到局部最小，得不到全局最小。

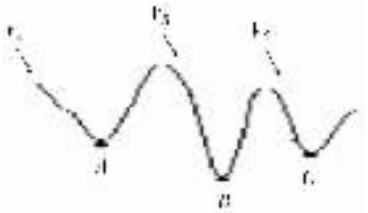

图 1 目标函数的局部极小和全局最小

2 动态隧道算法

文献[6]提出了动态隧道算法(Dynamic Tunneling Algorithm)，该算法就是用于在解全局最优化问题的时候使用动态优化系统和动态隧道系统来求解全局最优解。该算法由一系列的循环过程组成，在每个循环过程中，包括两个动态系统：动态优化系统和动态隧道系统。在动态优化系统中，找到一个局部的最小点，将这个局部最小点提供给动态隧道系统；在动态隧道系统中，以该局部最小点为初始值找到一个更小的能量盆地(Lower Energy Valley)，即找到提供给下一阶段的动态优化系统的初始值。这两个系统交替进行，直到在动态隧道系统中再也找不到更小的能量盆地为止。

下面对动态隧道系统的数学模型进行详细阐述。

全局最优化问题可以表示成如下的数学模型。

$$\min f(x)$$
$$\text{s.t. } g_j(x) < 0, j = 1,2,\cdots,m \qquad (2)$$

其中 x 是一个 n 维的向量，g 是 m 维的函数。在通常的情况下，$f(x)$ 称为目标能量函数，如果 $m = 0$ 该问题是一个无约束全局最优化问题。回到讨论由(2)式构成的能量问题求最小点的问题上，动态隧道系统采用以下的这个微分方程表示的动态系统来找到新的初始点。

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = -\frac{\dfrac{\partial f}{\partial x_i}}{[(x-x^*)^{\mathrm{T}}(x-x^*)]^{\lambda}} - \sum_{j=1}^{m} k_j f^*(g_j(x))\frac{\partial g_j(x)}{\partial x_i} - kf^*(\hat{f}(x))\frac{\partial f}{\partial x_i} \qquad (3)$$

$$f^*(z) = \begin{cases} z, & z > 0 \\ 0, & z < 0 \end{cases} \qquad (4)$$

$$k_j \geqslant \frac{\left\| \dfrac{\partial f}{\partial x} \right\|}{\left\| g_j(x)\dfrac{\partial g_j(x)}{\partial x} \right\|} (j = 1,2,\cdots,m) \qquad (5)$$

x^* 是由动态优化系统找到的局部最小点，$f^*(\cdot)$ 是(4)式定义的函数，k_j，k 是由(5)式来确定，$\hat{f}(x) = f(x) - f(x^*)$。可以注意到，当 x 在 x^* 的邻域内的时候，$\hat{f}(x) \geqslant 0$，因为 x^* 是 $f(x)$ 的局部最小点。

(3)式称为动态隧道系统，k 称为隧道惩罚。因 x^* 是平衡点，故 $(\partial f/\partial x) = 0$。在(3)式的第 1 项中引入分母 $[(x-x^*)^{\mathrm{T}}(x-x^*)]^{\lambda}$ 是为了去掉动态优化系统找到的平衡点，而第 3 项的引入是为了保证

找到一个新的初始点 $x^{(0)}$,$x^{(0)}$ 在目标能量函数的更低的能量盆地中,即 $f(x^{(0)}) \leqslant f(x^*)$,它是由(6)式求导得到。

$$k\int_0^{\hat{f}(x)} f^*(z)\,\mathrm{d}z \tag{6}$$

也就是说,用动态隧道系统来求解全局最优化问题时,在约束条件中需加上 $\hat{f}(x) = f(x) - f(x^*) \leqslant 0$。

动态隧道系统的平衡点与动态优化系统的平衡点满足 $\hat{f}(x) = f(x) - f(x^*) \leqslant 0$。可以证明动态系统(3)是一个平衡系统。因此从 x^* 偏移形成的动态系统(3)的动态流会汇聚到(3)式的平衡点,且动态隧道系统中每一次动态优化得到的局部最小值是满足以下条件

$$f(x^{1*}) \geqslant f(x^{2*}) \geqslant \cdots \tag{7}$$

即动态隧道算法中的每一次动态优化系统找到的局部最小值都小于或等于上一次找到的。

综上所述,动态隧道算法的总体流程可以归纳如下:设 x^* 是全局最优化问题(2)在上一次运行动态优化系统发现的最小值,如果在满足其约束条件的有效范围内存在一个点 $\hat{x}$ 满足 $f(\hat{x}) \leqslant f(x^*)$,那么可以通过动态隧道算法找到一个新的最小值 x^{**},满足 $f(x^{**}) \leqslant f(x^*)$。如果找到了一个最小点后,再也不能通过动态隧道算法找一个在更小能量盆地的点比新近找到的这个最小点更小,则这个新近找到的最小点就是全局最小点。

3 基于动态隧道系统的 K-means 聚类算法

K-means 聚类算法在迭代过程中由于聚类中心不再改变,而认为目标函数已达到全局最小值点而终止算法,实际上,此时往往找到的是局部极小点,这是由于算法本身没有有效的改进措施使得极小值点可以跳出其极小值区域,转移到一个新的区域继续迭代寻优而造成的。动态隧道算法是在用动态优化过程寻找到一个极小值点后,再利用动态隧道过程以该极小值点作为初始点寻找一个更小的能量盆地,反复重复这两个过程直到达到全局最小点。将其中的动态隧道过程引入到 K-means 聚类算法中,在 K-means 聚类算法找到一个极小值点后,再利用动态隧道系统来开辟一个新的搜索方向,使其能流向能量更小的区域中的一个点,将其提供给 K-means 聚类算法继续迭代优化此区域以达到局部最小。反复此过程,直到动态隧道系统无法再找到更小的能量盆地为止。

采用以下动态系统作为动态隧道系统

$$\mathrm{d}V_j/\mathrm{d}t = \rho(V_j - V_j^*)^{1/3} \tag{8}$$

(8)式中,ρ 表示学习强度,V_j^* 是相对于 V_j 上一次的局部最小值。可以看出,对这个动态隧道系统,V^* 就是这个系统的平衡点。$V = V^* + \varepsilon$,其中 ε 为一定的步长,这样每一个时间步都会引导聚类中心值沿着隧道的某个搜索方向前进。每改变一个聚类中心,对应聚类中心值空间中的一个新的点,所以要重新根据(1)式计算 J_c,如果 $J_{cV_j} \leqslant J_{cV_j^*}$,则动态隧道过程结束,找到了 K-means 聚类算法的新起点。整个算法停止于由动态隧道过程无法再寻找到新的更小能量盆地,即不能找到新的起始点作为 K-means 聚类算法的起始点。那么上一次 K-means 聚类算法找到的最小值点就是全局最小值点。

基于动态隧道系统的 K-means 聚类算法的算法流程如下。

Step1　给定数据规模为 n 的数据集,令 $I = 1$,选取 k 个初始聚类中心 $V_j(I)$,$j = 1,2,3,\cdots,k$;

Step2　计算每个数据对象与聚类中心的距离,$D(x_i, V_j(I))\ i = 1,2,3,\cdots,n, j = 1,2,3,\cdots,k$,如果满足 $D(x_i, D_k(I)) = \min\{D(x_i, V_j(I)), j = 1,2,3,\cdots,n\}$,则 $x_i \in w_k$;

Step3　按照(1)式计算误差平方和准则函数 J_c;

Step4　若 $|J_c(I) - J_c(I-1)| < \xi$,则转到 Step 5;否则 $I = I + 1$,计算 k 个新的聚类中心,$V_j(I) = \frac{1}{n}\sum_{i=1}^{n_j} x_i^{(j)}$,$j = 1,2,3,\cdots,k$,返回 Step 2;

Step5　$V \leftarrow V^* + \varepsilon$;

for $t = 1$ to N // 遍历每个时间步

按照(8)式对一个聚类中心 V_j 进行积分得到 $V_{j(t)}$;

计算 $J_c(v_{j(t)})$;

计算 $\delta = J_c(V_{j(t)}) - J_c(V^*)$;

if $\delta < 0$,则找到新的 K-means 聚类算法起始点,退出动态隧道过程,$I = 1$,返回 Step 2;

endfor

Step6　若 $\delta \geqslant 0$,则结束整个算法,V^* 为全局最小点。

4 实验结果

为了检验本文提出的基于动态隧道系统的

K-means聚类算法的有效性,笔者将其与K-means聚类算法应用在一个数据规模为1100的二维仿真数据集上,进行对比实验,比较新算法与原算法的聚类效果。

数据集如图2(a)所示,包括4个大小相当的球形簇,且簇与簇之间区别明显。这种数据集分布比较理想。算法迭代收敛的条件ξ设置为1E-6。

图2 仿真数据分布及实验结果

再考察K-means聚类算法改进前后产生聚类结果的稳定性及准确性。实验分别用K-means聚类算法(指定$K=4$)和基于动态系统的K-means聚类算法对数据集分别进行10次聚类,聚类中心在图中用黑色小矩型代表,以它们为中心的类也分别用虚线圆圈表示,并配有阿拉伯数字标注的类标号。

两种算法对仿真数据集的处理结果见表1,表中第1列为实验次数,第2列为使用K-means聚类算法得到的准则函数J_c值,第4列为基于动态隧道系统的K-means聚类算法得到的准则函数J_c值。第3列和第5列表示算法能否正确聚类。

从表1中可以看出,对于这种分布比较理想的数据集,应用K-means聚类算法聚类虽然可以得到正确的聚类结果,但是并不稳定,聚类中心变化非常明显,聚类效果不好的情况存在的概率较大。实验正确率为60%,10次实验只有6次可以正确聚类,另外4次都出现了类似图1(b)陷入局部极小的情况,聚类中心为:(0.775,1.282),(0.394,0.578),(1.285,0.541),(0.315,0.432)。其J_c值远大于正确聚类时的值,聚类效果相当不理想。而采用基于动态隧道系统的K-means聚类算法,由于能够通过在每次迭代中找到更小的能量盆地,实验正确率为100%,10次实验中每次都能得到正确的聚类结果,如图1(c)所示,聚类中心为:(0.421,1.210),(0.341,0.532),(1.070,1.284),(1.291,0.556)。显然,改进算法能够保证每次收敛的结果都是正确一致的,不仅在聚类效果上并且在稳定性上都远远优于K-means聚类算法,在理论上和实践上是可行的。

表1 两种算法的J_c值比较

实验次数	K-means聚类算法	能否正确聚类	基于动态隧道系统的K-means聚类算法	能否正确聚类
1	17.231	是	17.231	是
2	68.297	否	17.231	是
3	90.998	否	17.231	是
4	17.231	是	17.231	是
5	17.231	是	17.231	是
6	17.231	是	17.231	是
7	129.106	否	17.231	是
8	16.121	是	17.231	是
9	68.518	否	17.231	是
10	17.231	是	17.231	是

5 结论

为了有效避免K-means聚类算法易陷入局部极小值的缺陷,本文提出了一种基于动态隧道系统的K-means聚类算法,该算法以K-means聚类算法所得到的局部极小值作为动态隧道过程的初始点,寻找出一个更小的能量盆地,再提交给K-means聚类算法来加以优化,整个过程重复执行,直到无法利用动态隧道过程再寻找到更小能量盆地为止,此时得到全局最小值点。仿真实验结果表明,该算法是正确可行的。

参考文献:

[1] Fayyad U, Reina C, Bradley P S. Initialization of iterative refinement clustering algorithms [C]. MenloPark: AAAI, 1998, 194-198.

[2] 吕佳. 可能性C-Means聚类算法的仿真实验[J]. 重庆师范大学学报(自然科学版),2005,22(3):129-132.

[3] 吕佳. 核聚类算法及其在模式识别中的应用[J]. 重庆师范大学学报(自然科学版),2006,23(1):22-24.

[4] Krishma K, Murty M N. Genetic K-means algorithm [J]. IEEE Transactions on System, Man, and Cybernetics, Part B. 1999. 29(3):433-439.

[5] 吕佳. 基于免疫聚类的Web日志挖掘[J]. 重庆师范大学学报(自然科学版),2007,24(2):32-35.

[6] Yong Yao. Dynamic tunneling algorithm for global optimization[J]. IEEE Transactions on Systems, Man and Cybernetics,1989,19(5):1222-1230.

[7] 刘贞,张小真. 基于最小聚类单元的商圈聚类方法研究[J]. 西南师范大学学报(自然科学版),2004,29(6):949-952.

[8] Han J, Kamber M. 数据挖掘概念与技术[M]. 范明,孟小峰译. 北京 机械工业出版社,2002.

Research into K-means Clustering Algorithm Based on Dynamic Tunneling System

LÜ Jia

(College of Mathematics and Computer Science, Chongqing Key Lab. of Operations Research and System Engineering, Chongqing Normal University, Chongqing 400047, China)

Abstract: K-means clustering algorithm itself is a global optimization problem, whose objective function has multiple local minima and only one global minimum. The algorithm is apt to fall into local minimum points through its iteration process. Dynamic tunneling algorithm is especially developed for the global optimization problem. On the basis of the local minimum got by dynamic optimization process, dynamic tunneling process makes use of tunneling to skip the local minimum point and search a lower energy valley that in essence means a lower value point than the local minimum point, thus a new initial point obtained can further be transferred to dynamic optimization process to be optimized. The use of the validity of dynamic tunneling algorithm, dynamic tunneling process of the algorithm is introduced into K-means clustering algorithm, and a novel K-means clustering algorithm based on dynamic tunneling system is presented in this paper. On the basis of local minimum got by K-means clustering algorithm, dynamic tunneling process is used to search a lower energy valley, then the value is submitted to K-means clustering algorithm for iterative optimization. The process is unceasingly repeated until the global minimum point is found. Both the theory analysis and simulation experiment results show that the presented algorithm in this paper is superior to K-means clustering algorithm.

Key words: K-means clustering algorithm; global optimization; objective function; dynamic tunneling system; energy valley

(责任编辑 游中胜)

(上接第 72 页)

A M/M/1 Queuing Model with Variable Input Rate

TAI Wen-zhi, GAO Shi-ze

(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: The M/M/1 queuing theory of variable input rate is an important queuing model. Customers are frequently seen in our daily life hesitating about joining the long queue when they find so many people waiting in the line in front of the service desk. Generally speaking, the probability for customers joining the queue changes with the length of it. This paper, by discussing the M/M/1 queuing model when the customers arrived entering this queuing system by the probability of $\alpha_k = \frac{1}{\beta k + 1}$, aims at getting the stable distribution and the related indexes of this modals, such as the average input rate of customers, the average intensity of the service of the system, the average number of the customers, the average number of the customers of the system, the loss probability of the system, the probability of customers getting into the system and receiving service, the average number of customers getting into the system in a given time, the average number of customers lost in a given time ect. Thus it extends the application of the results in reference [1].

Key words: queuing system; variable input rate; sojourn time; waiting time

(责任编辑 游中胜)